**西南财经大学**

 Southwestern University of

Finance and Economics

R语言课程作业

**学生姓名： 杨双杰**

**所在学院： 会计学院**

**专 业： 财务管理**

**学 号： 2161202Z6024**

**任课教师： 李 伊**

2018年1月

一、数据说明



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量选取列表 | | | | | | | | | |
| 被解释变量 |  | | 变量说明 | | | | | | |
| Opinion\_dum | 审计意见，虚拟变量，事务所出具非标准审计意见则取值1，否则为0 | | | | | | | | |
| lnfee | 审计费用，取对数 | | | | | | | | |
| 解释变量 |  |  | |  |  |  |  |  |  |
| downturn | 经济周期，虚拟变量，审计当年处于经济衰退期取值1，否则为0 | | | | | | | | |
| 控制变量 |  |  | |  |  |  |  |  |  |
| lnasset | 企业规模，年末总资产对数 | | | | | | | | |
| roa | 盈利能力，资产收益率 | | | | | | | | |
| lev | 财务杠杆，资产负债率 | | | | | | | | |
| loss | 盈亏状况，虚拟变量，公司盈利取值1，亏损为0 | | | | | | | | |
| growth | 公司成长性，营业收入增长率 | | | | | | | | |
| ndts | 非债务税盾，固定资产折旧、无形资产摊销等资产折耗占总资产比率 | | | | | | | | |
| soe | 控制人性质，虚拟变量，公司实际控制人为国有，取值为1，否则为0 | | | | | | | | |
| sanction | 监管力度，每年被证监会处罚的事务所数量 | | | | | | | | |
| complex | 审计复杂程度，公司存货与应收账款占总资产的比例 | | | | | | | | |
| peratio | 投资需求，市盈率 | | | | | | | | |
| ind | 行业，虚拟变量，若为所属行业，取值为1，否则为0 | | | | | | | | |

这个模型用于研究经济周期X（downturn）对审计质量Y（opinion\_dum和lnfee）具有怎么的影响关系，此外模型中除了行业（ind）、年度（year）外，还包含其他10个控制变量（controls），各变量具体说明如下表。

为了得到各个变量的数据，从CSMAR，DIB数据库里一共下载了近20份原始数据，现在利用R对这些数据进行清洗最后合并为一份最终数据。具体流程主要包括：①清洗各零散的控制变量，然后合并为一份“control\_data”。②清洗解释变量得到含有经济周期虚拟变量downturn的数据“GDP\_grate”，清洗被解释变量得到数据“audit”。③最后将所有变量合并为一份“final\_data”，并结合前面清洗的数据绘制3张图。

1. 清洗流程
2. 清洗控制变量

首先设置项目工作路径，加载清洗所需要的R包。

#Tidy Control Variables ------------------------------------------------

setwd("./control")

library(tidyverse)

library(lubridate)

1. 第一份数据读取主要包含资产负债表变量的balance.csv文件，并自定义变量名，筛选出合并报表数据（A）以及年度数据（月份“12”），并删除不需要的四个变量（typrept,fasset,fadisposal,iasset），数据存为balance。然后让年度变量year只显示年度，同时由于股票代码stkcd前面的0被省去，此处统一股票代码的数据格式为6位数，以便后续数据合并。

# the first data

col\_names <- c("stkcd", "year", "typrept", "recei",

"inv", "fasset", "fadisposal", "iasset", "assets")

balance <- read\_csv("balance.csv",col\_names = col\_names) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept,-fasset,-fadisposal,-iasset)

balance$year <- year(ymd(balance$year))

balance$stkcd <- sprintf("%06d",as.numeric(balance$stkcd))

2.第二份数据读入grevenue.csv文件，并自定义变量名，筛选合并报表和年度数据，之后删除不需要的变量，对年度year和股票代码stkcd的处理同上。

#the second data

growth <- read\_csv("grevenue.csv",col\_names = c(

"stkcd", "year", "typrept", "ind", "grev", "susgrate")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept,-susgrate)

growth$year <- year(ymd(growth$year))

growth$stkcd <- sprintf("%06d",as.numeric(growth$stkcd))

3.第三份数据处理过程同前。

#the third data

lev <- read\_csv("LEV.csv",col\_names = c(

"stkcd", "year", "typrept", "ind", "lev")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept)

lev$year <- year(ymd(lev$year))

lev$stkcd <- sprintf("%06d",as.numeric(lev$stkcd))

4.第四份数据处理过程同前。

#the fourth data

depamor <- read\_csv("depamor.csv",col\_names = c(

"stkcd", "year", "typrept", "ind", "depamor")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept)

depamor$year <- year(ymd(depamor$year))

depamor$stkcd <- sprintf("%06d",as.numeric(depamor$stkcd))

5.第五份数据处理过程同前。

#the fifth data

netincome <- read\_csv("NI.csv",col\_names = c(

"stkcd", "year", "typrept", "ni")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept)

netincome$year <- year(ymd(netincome$year))

netincome$stkcd <- sprintf("%06d",as.numeric(netincome$stkcd))

6.第六份数据处理过程同前。

#the sixth data

roa <- read\_csv("ROAE.csv",col\_names = c(

"stkcd", "year", "typrept", "ind", "roa","roe","roic")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept,-roe,-roic)

roa$year <- year(ymd(roa$year))

roa$stkcd <- sprintf("%06d",as.numeric(roa$stkcd))

7.第七份数据处理过程同前。

#the seventh data

roegrate <- read\_csv("roegrate.csv",col\_names = c(

"stkcd", "year", "typrept", "ind", "roegrate")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept)

roegrate$year <- year(ymd(roegrate$year))

roegrate$stkcd <- sprintf("%06d",as.numeric(roegrate$stkcd))

8.第八份数据不含子公司数据，所以不需要删选出合并报表数据（A），其余过程同前。

#the eighth data

peratio <- read\_csv("PEratio.csv",col\_names = c(

"stkcd", "year", "ind", "peratio")) %>%

filter(str\_sub(year,6,7)=="12")

peratio$year <- year(ymd(peratio$year))

peratio$stkcd <- sprintf("%06d",as.numeric(peratio$stkcd))

9.第九份数据为对审计师的行政处罚（saction）数据，这里只需读入并删除一个不需要的数据（discipline）。

#the nineth data

saction <- read\_csv("saction.csv") %>%

select(-discipline)

10.第十和第十一份数据共同得出需要的公司性质变量（soe），由两份不同来源的数据经过清洗合并得出。首先读入DIB来源的数据，由于含有中文，读入时指定编码格式为GB18030，筛选所需的1，4，5列的变量，并重命名，再生成一个缺失值变量soe用于根据公司性质type变量是否含有关键字段而分别赋予0和1，即二值变量。第十一份数据处理过程与此类似。最后将两份soe数据进行full\_join,再利用unique去重。

#the tenth and eleventh data

soe1 <- read\_csv("境内公司基本信息-公司性质-DIB.csv",

locale = locale(encoding = "gb18030")) %>%

select(c(1,4,5)) %>%

rename(

stkcd = 证券代码,

type = 公司性质,

controller = 实际控制人

) %>%

mutate(soe=NA)

soe1$stkcd <- str\_sub(soe1$stkcd,1,6)

soe1$soe[soe1$type=="国有企业"] <- 1

soe1$soe[soe1$type=="其他" & str\_detect(

soe1$controller,"国有资产监督管理")] <- 1

soe1$soe[is.na(soe1$soe)] <- 0

soe1 <- soe1 %>%

select(stkcd,soe)

soe2 <- read\_csv("上市公司实际控制人.csv",

locale = locale(encoding = "gb18030")) %>%

select(c(1,4)) %>%

rename(

stkcd = 代码,

contr\_type = 控制人类型

) %>%

mutate(soe=NA)

soe2$stkcd <- str\_sub(soe2$stkcd,1,6)

soe2$soe[str\_detect(

soe2$contr\_type,"国资委|国有企业|中央|政府")] <- 1

soe2$soe[is.na(soe2$soe)] <- 0

soe <- soe2 %>%

select(-contr\_type) %>%

full\_join(soe1) %>%

unique()

1. 第十二份和第十三份数据分别为公司ST年份和IPO年份数据，这里处理过程依然与前述相同。只是进行了保存输出到final文件夹下。

#the twelveth and thirteenth data

ST <- read\_csv("ST.csv",col\_names = c("stkcd", "year", "typrept")) %>%

filter(typrept=="A",str\_sub(year,6,7)=="12") %>%

select(-typrept) %>%

mutate(st=1)

ST$year <- year(ymd(ST$year))

ST$stkcd <- sprintf("%06d",as.numeric(ST$stkcd))

write\_csv(ST,"../final/ST.csv")

IPO <- read\_csv("IPO.csv",skip = 1,col\_names = c(

"stkcd", "initial", "ipoyear", "listyear")) %>%

filter(initial=="A") %>%

select(stkcd,ipoyear)

IPO$ipoyear <- year(ymd(IPO$ipoyear))

IPO$stkcd <- sprintf("%06d",as.numeric(IPO$stkcd))

write\_csv(IPO,"../final/IPO.csv")

12.这里将前述所有清洗的控制变量利用reduce函数进行full合并，并利用其他变量生成新的所需变量lnassets、ndts、complex和缺失值变量loss，然后根据净利润ni大于还是小于0分别对loss赋值0或1。最后保存输出为control\_data.csv。

#integrated controls

control\_data <- list(balance,depamor,growth,lev,

netincome,roa,roegrate,peratio,soe,saction) %>%

reduce(full\_join) %>%

mutate(

lnassets=log(as.numeric(.$assets)),

loss = NA,

ndts = as.numeric(.$depamor)/as.numeric(.$assets),

complex=(

(as.numeric(.$inv)+as.numeric(.$recei))/as.numeric(.$assets)

) %>%

select(-recei,-inv,-assets,-depamor)

control\_data$loss[control\_data$ni>=0] <- 1

control\_data$loss[control\_data$ni<0] <- 0

control\_data <- control\_data %>%

select(-ni)

write\_csv(control\_data,"../final/control\_data.csv")

1. 清洗解释变量和被解释变量

先将工作路径切换到XY文件夹。

#Tidy XY -------------------------------------------------------------

setwd("../XY")

1. 处理X。读入GDP数据，根据各年的GDP指数计算GDP增长率（gdpgrate），然后根据本年增长率相较于去年是否下降生成虚拟变量downtown。然后选取所需变量存为GDP\_cycle数据，并输出为X.csv。

GDP <- read\_csv("GDP.csv",col\_names = c("year","gdpindex"),

locale=locale(encoding = "gb18030"),skip = 7) %>%

select(-3) %>%

mutate(gdpgrate=(as.numeric(.$gdpindex)-100)/100 ,

downturn = NA)

GDP$downturn[GDP$gdpgrate<lag(GDP$gdpgrate)] <- 1

GDP$downturn[is.na(GDP$downturn)] <- 0

GDP\_cycle <- select(GDP,'year',"downturn")

write\_csv(GDP\_cycle,"../final/X.csv")

1. 处理Y。读入数据后，选取所需列并重命名，再过滤掉非年度数据，生成审计费用的对数（lnfees）。之后根据审计意见opinon变量是否为“标准无保留意见”对opinion\_dum分别赋值0和1。最后删除第3和4列不需要的变量输出为Y.csv。

audit <- read\_csv("audit.csv",locale=locale(encoding = "gb18030")) %>%

select(1,3,5,13) %>%

rename(

stkcd = 证券代码,

year = 会计截止日期,

opinion = 审计意见类型,

fees = 审计费用合计

) %>%

filter(str\_sub(year,6,7)=="12") %>%

mutate(

lnfees = log(as.numeric(fees)),

opinion\_dum = NA

)

audit$year <- year(ymd(audit$year))

audit$stkcd <- sprintf("%06d",as.numeric(audit$stkcd))

audit$opinion\_dum[audit$opinion=="标准无保留意见"] <- 0

audit$opinion\_dum[is.na(audit$opinion\_dum)] <- 1

audit <- select(audit,-3,-4)

write\_csv(audit,"../final/Y.csv")

1. 将所有变量合并清洗为最终数据

#Tidy Final Data --------------------------------------------------------

setwd("../final")

这里先读取行业（ind）数据，这份数据较全，有利于减少最终数据中行业的缺失值。过程同样如前大多数处理，读入-命名-选取合并报表数据（A）-年度数据（12）-删除不需要的报告类型变量（-typre）。

ind <- read\_csv("ind.csv",col\_names = c(

"stkcd","year","typre","ind"),skip=1) %>%

filter(typre=="A",str\_sub(year,6,7)=="12") %>%

select(-typre)

ind$year <- year(ymd(ind$year))

ind$stkcd <- sprintf("%06d",as.numeric(ind$stkcd))

将前面各步骤处理的数据作为列表传入reduce进行full\_join操作，然后分别剔除ST年和IPO年度数据，再剔除金融业（行业代码J开头，股票代码2或9开头），最后剔除资产负债率（lev）大于1的数据，最终再剔除缺失值和重复值。

final\_data <- list(audit,GDP\_cycle,control\_data,ST,IPO,ind) %>%

reduce(full\_join) %>%

filter(st!=1|ipoyear!=year) %>%

filter(!str\_detect(ind,"^J")) %>%

filter(!str\_detect(stkcd,"^(2|9)")) %>%

filter(lev<1) %>%

select(-st,-ipoyear) %>%

drop\_na() %>%

unique()

write\_csv(final\_data,"finaldata.csv")

三、利用清洗的数据作图

（一）第一张图为不同经济周期对数审计费用的统计差异

#picture 1

final\_data %>%

ggplot() +

stat\_summary(aes(x=downturn,y=lnfees,color=downturn),

show.legend = FALSE,

fun.ymin = min,

fun.ymax = max,

fun.y = mean) +

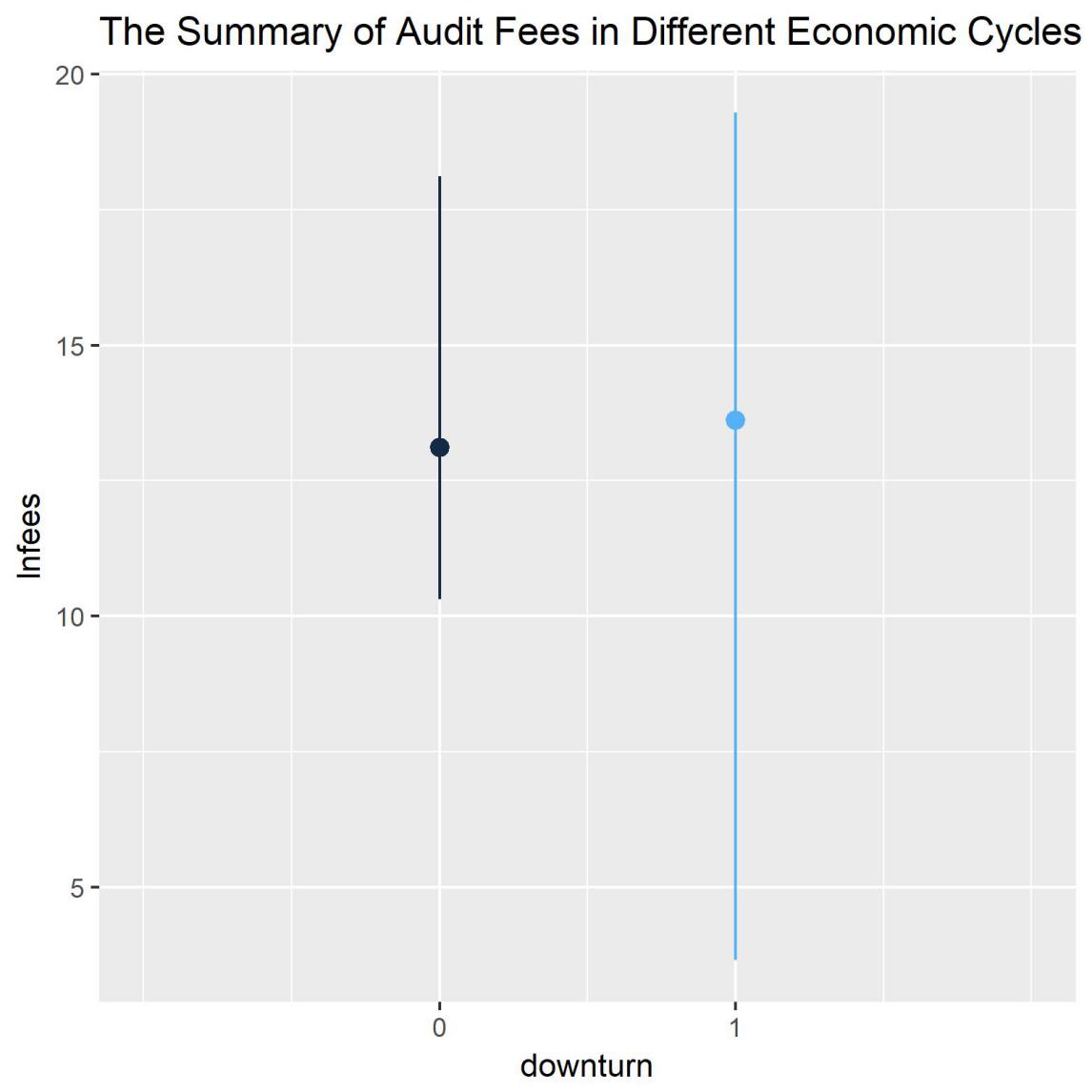
labs(title = paste(

"The Summary of Audit Fees in Different Economice Cycles"

)) +

scale\_x\_continuous(limits=c(-1,2),breaks = seq(0,1))

ggsave("diff\_lnfees.jpeg")



1. 第二张图为GDP增长率和审计费用增长率的变化图，这里先根据年度分组计算对数审计费用的平均值，剔除平均后各年的重复值，进而计算对数审计费用的增长率，最后与处于不同表格的GDP增长率进行left\_join操作。

#picture 2

GDP\_lnfees <- audit %>%

select(stkcd,year,lnfees) %>%

group\_by(year) %>%

mutate(mean\_lnfees = mean(lnfees,na.rm=TRUE)) %>%

select(year,mean\_lnfees) %>%

unique() %>%

ungroup() %>%

mutate(

lnfeesgrate = (mean\_lnfees-lag(mean\_lnfees))/lag(mean\_lnfees)

) %>%

left\_join(GDP) %>%

select(year,lnfeesgrate,gdpgrate) %>%

drop\_na()

ggplot(data=GDP\_lnfees) +

geom\_line(mapping = aes(x=year,y=gdpgrate),color="red") +

scale\_x\_continuous(breaks = GDP\_lnfees$year) +

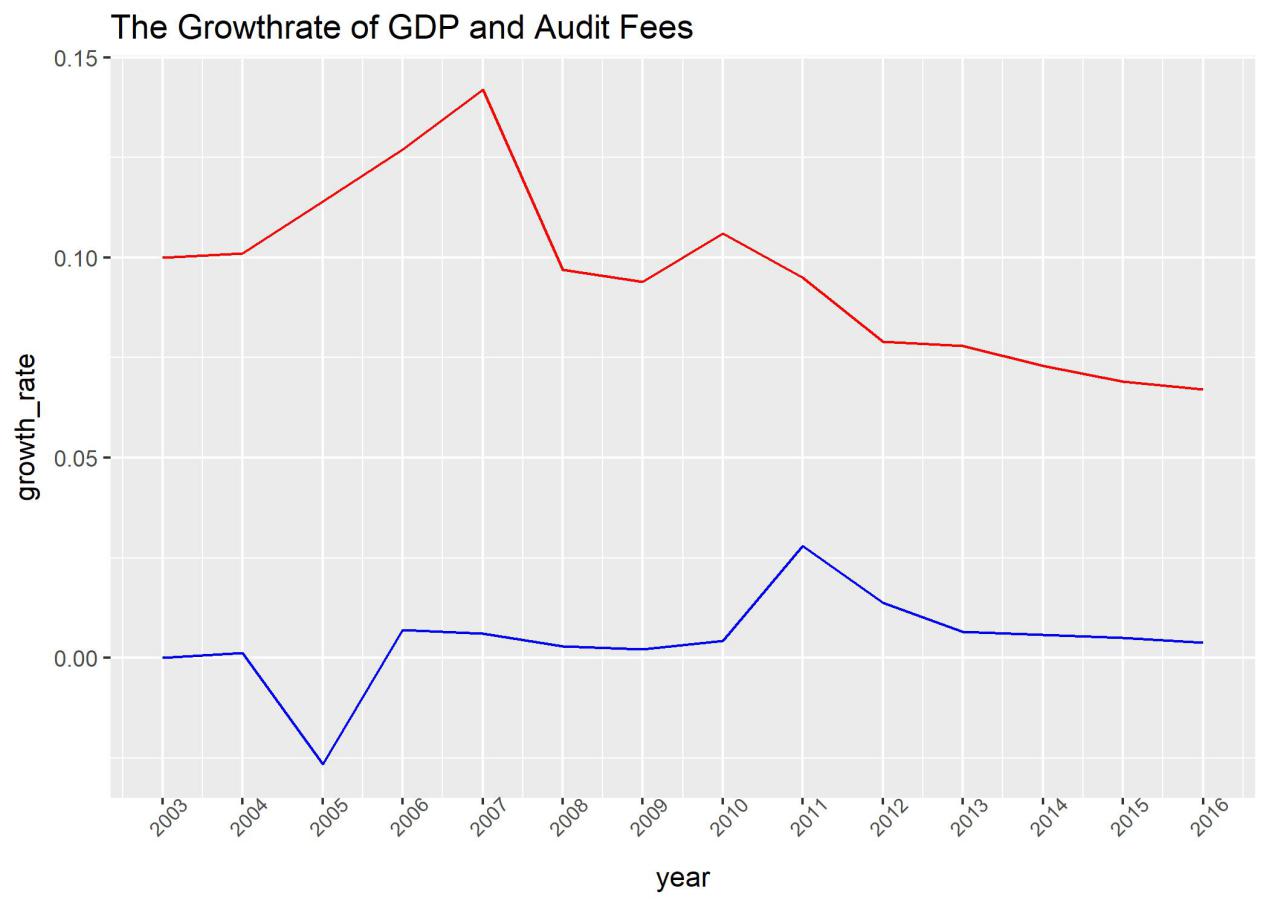
theme(axis.text.x = element\_text(angle=45,size=8)) +

geom\_line(mapping=aes(x=year,y=lnfeesgrate),color="blue") +

labs(y="growth\_rate",

title="The Growthrate of GDP and Audit Fees ")

ggsave("growth\_rate.jpeg")



（三）第三张图为不同经济周期审计意见的分布图。

#picture 3

final\_data %>%

ggplot() +

geom\_bar(aes(x=opinion\_dum,fill=downturn),show.legend = FALSE) +

facet\_wrap(~downturn) +

labs(

title="The Distribution of Auditing

Opinions in Different Cycles"

) +

scale\_x\_continuous(breaks=seq(0,1,1))

ggsave("opinion\_distribution.jpeg")

